Software code for performing instrumental variable analyses for Mendelian randomization investigations

maintained by Stephen Burgess

November 9, 2015

This is a non-traditional publication to provide software code for the Mendelian randomization community in a single document. It will be updated when necessary as new methods are developed. Hopefully, this will become a collaborative resource than can be authored by the community rather than a single-author manuscript. However, Stephen Burgess retains the prerogative to exert editorial control.

Currently, it mostly contains R code. If someone wants to write Stata code or code for any other software package, this could be included.

Contributors:

- Stephen Burgess (sb452@medschl.cam.ac.uk)
- James Staley (jrs95@medschl.cam.ac.uk)
- Tom Palmer
- Jack Bowden

Contents

1 Introduction and notation

```
### Dimensions
N # sample size
K # number of genetic variants

### Individual-level data
g # genetic variant(s), matrix dimension N x K
x # risk factor/exposure, vector length N
y # outcome, vector length N

###### Summarized data
bx   # genetic associations with exposure, vector length K
by   # genetic associations with outcome, vector length K
bxse # standard errors of genetic associations with exposure
byse # standard errors of genetic associations with outcome
```

This document describes methods for causal estimation using Mendelian randomization. Little attention is paid as to the assumptions for Mendelian randomization, or the interpretation of estimates from these methods. See Hernán and Robins [2006], Glymour et al. [2012], VanderWeele et al. [2014] and Burgess et al. [2015a] for some critical comments.

2 Mendelian randomization analysis with individual-level data

This section on standard Mendelian randomization methods with individual-level data in a single dataset is based on Burgess et al. [2015e], which in turn is based on Chapter 4 of Burgess and Thompson [2015b]. We consider in turn the ratio of coefficients method, two-stage methods, likelihood-based methods, and semi-parametric methods.

2.1 Ratio of coefficients (Wald) method – single instrument

The ratio of coefficients method, or the Wald method is the simplest way of estimating the causal effect of the risk factor on the outcome (original paper [Wald, 1940]). The ratio method uses a single instrumental variable (IV), which can be a single SNP or an allele score (see Burgess and Thompson [2013] for background on allele scores).

```
## A. Ratio estimate (continuous outcome)

bx   = lm(x~g)$coef[2]
bxse = summary(lm(x~g))$coef[2,2]
by   = lm(y~g)$coef[2]
byse = summary(lm(y~g))$coef[2,2]

beta_ratio = by/bx
```

See Greenland [2000] or Martens et al. [2006] for an introduction to instrumental variable methods and causal estimation, or Lawlor et al. [2008] for a specific Mendelian randomization perspective.

```
## B. Asymptotic standard error (poor with weak instruments)

# 1. Delta method approximation (summarized data)
se_ratio_approx = byse/bx
        # first order approximation
se_ratio_approx = sqrt(byse^2/bx^2 + by^2*bxse^2/bx^4 - 2*theta*by/bx^3)
        # second order approximation
        # theta is correlation between numerator
        #   and denominator in ratio estimate
        # theta = 0 in a two-sample setting

# 2. Two-stage least squares method for standard error (individual-level
    data)
library(sem)
se_tsls = sqrt(tsls(y, cbind(x, rep(1,N)), cbind(g, rep(1,N)),
    w=rep(1,N))$V[1,1])

library(ivpack)
ivmodel = ivreg(y~x|g, x=TRUE)
summary(ivmodel)$coef[2,2]
```

Asymptotic confidence intervals tend to be too narrow when the instrument is weak (that is, it does not explain much of the variance of the risk factor in the population), as the distribution of the ratio estimator is non-normal with heavy tails. Hence methods have been developed that provide confidence intervals without making the assumption that the ratio estimator has a normal distribution.

```
## C. Valid confidence intervals with weak instruments

# 1. Fieller's theorem
```

```
  bx   = lm(x~g)$coef[2]
  bxse = summary(lm(x~g))$coef[2,2]
  by   = lm(y~g)$coef[2]
  byse = summary(lm(y~g))$coef[2,2]

  f0 = by^2 - qt(0.975, N)^2 * byse^2
  f1 = bx^2 - qt(0.975, N)^2 * bxse^2
  f2 = by*bx
   D = f2^2 - f0*f1

  if(D>0) {
    r1 = (f2-sqrt(D))/f1
    r2 = (f2+sqrt(D))/f1
if(f1>0) { cat("Confidence interval is a closed interval [a,b]: \n a=",
    r1, ", b=", r2, sep="") }
if(f1<0) { cat("Confidence interval is the union of two open intervals
    (-Inf, a], [b, +Inf): \n a=", r2, ", b=", r1, sep="") }
         }
if(D<0|D==0) { cat("No finite confidence interval exists other than the
    entire real line.") }

# 2. Anderson--Rubin
library(ivpack)
ivmodel = ivreg(y~x|g, x=TRUE)
anderson.rubin.ci(ivmodel)
        # As with Fieller's theorem, interval may be a closed interval,
            the union of two open intervals, or undefined
```

A reference for Fieller's theorem is Buonaccorsi [2005] (original reference is Fieller [1954], a web-based tool is available at `https://sb452.shinyapps.io/fieller`. A reference for the Anderson–Rubin method is Mikusheva [2010] (original reference is Anderson and Rubin [1949]).

```
## D. Binary outcome, logistic-linear model (assuming case--control data)

bx   = lm(x[y==0]~g[y==0])$coef[2]
bxse = summary(lm(x[y==0]~g[y==0]))$coef[2,2]
by   = glm(y~g, family=binomial)$coef[2]
byse = summary(glm(y~g, family=binomial))$coef[2,2]

beta_ratio     = by/bx
se_ratio_approx = byse/bx # ...and so on.
```

With a binary outcome in a case–control setting, genetic associations with the risk factor should be estimated in control participants only (see Didelez and Sheehan [2007]). This is for three main reasons: to avoid reverse causation, to avoid biases due to outcome-dependent sampling, and because the controls are a more representative sample of the population as a whole. If pre-disease measurements of the risk factor are available for cases, and the prevalence of the disease is available (such as in a

nested case–control setting), then associations may be estimated in the case–control provided that participants are weighted so that the case-control sample represents the underlying population [Bowden and Vansteelandt, 2011].

There are some technical issues relating to the ratio estimate with a binary outcome and a logistic regression model due to the non-collapsibility of odds ratios [Greenland et al., 1999], but it is a consistent estimator under the null [Vansteelandt et al., 2011]. Some further references for instrumental variable analysis with a binary outcome are Palmer et al. [2011] and Clarke and Windmeijer [2012].

Much of the code for a continuous outcome, including Fieller's theorem, can be used with a binary outcome.

2.2 Two-stage methods

Two-stage methods are often implemented in practice in a single step, but can be estimated using two separate regression models. The first-stage model consists of a regression of the risk factor on the instrumental variables. The second-stage model consists of a regression of the outcome on the risk factor. If the model is calculated using these two stages (a sequential regression approach), then standard errors from the second-stage regression model will be underestimated as they will fail to account for uncertainty in the first-stage regression. However, uncertainty in the first-stage model may be small and can often be neglected.

```
## A. Continuous outcome -- two-stage least squares

library(sem)
beta_tsls = tsls(y, cbind(x, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$coef[1]
se_tsls  = sqrt(tsls(y, cbind(x, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$V[1,1])

library(ivpack)
ivmodel = ivreg(y~x|g, x=TRUE)
summary(ivmodel)
beta_tsls = ivreg(y~x|g, x=TRUE)$coef[2]
se_tsls  = summary(ivreg(y~x|g, x=TRUE))$coef[2,2]
```

The same point estimate (although not the standard error) can be obtained using sequential regression:

```
beta_seqreg_tsls = lm(y~lm(x~g)$fitted)$coef[2]

## B. Binary outcome, logistic-linear model (assuming case--control data)

# 1. Two-stage predictor substitution.
g0       = g[y==0]
tsps.glm = glm(y~predict(lm(x[y==0]~g0), newdata=list(g0=g)),
   family=binomial)
beta_tsps = tsps.glm$coef[2]
se_tsps  = summary(tsps.glm)$coef[2,2]
```

This is also known as two-stage predictor substitution as the fitted values from the first-stage regression are plugged into the second-stage regression (as in sequential regression) [Cai et al., 2011]. The estimand with a logistic regression model in the second stage is a population-averaged causal odds ratio [Burgess and CCGC, 2012]. This represents the ratio in odds for a intervention in the population distribution of the risk factor averaged over the population. It differs from the subject-specific (or individual) odds ratio even if this is constant due to non-collapsibility. It still provides a valid test of the causal null hypothesis, and is it consistent under the null [Vansteelandt et al., 2011].

```
# 2. Two-stage residual inclusion.
```

```
g0       = g[y==0]
pred     = predict(lm(x[y==0]~g0), newdata=list(g0=g))
resid    = x-predict(lm(x[y==0]~g0), newdata=list(g0=g))
tsri.glm = glm(y~pred+resid, family=binomial)
beta_tsri = tsri.glm$coef[2]
se_tsri = summary(tsps.glm)$coef[2,2]
```

An alternative approach is two-stage residual inclusion, in which the fitted values and residuals from the first-stage regression are both included in the second-stage regression model [Terza et al., 2008]. This is also known as the control function approach [Nagelkerke et al., 2000]. Inclusion of the residual term means that estimates from the second-stage regression model are closer to subject-specific odds ratios [Palmer et al., 2008]. However, there is no good reason why this adjustment should be made, and estimates from the two-stage residual inclusion method may even be biased under the null [Vansteelandt et al., 2011]. Hence the two-stage predictor substitution method should be preferred in practice.

2.3 Likelihood-based methods

Not aware of a generic implementation of likelihood-based methods (such as limited information maximum likelihood, LIML) in R. The LIML approach can be implemented in Stata:

```
ivreg2 y (x = g1), liml      * for one variant
ivreg2 y (x = g1 g2 g3), liml * for multiple variants.
```

2.4 Semi-parametric methods

Semi-parametric methods are those that make some parametric assumptions, but fewer assumptions than a fully-parametric method. For instance, the structural model relating the risk factor to the outcome may be specified (for instance, as a linear model), but the distribution of the errors in this model are not specified. The category of semi-parametric methods includes the generalized method of moments (GMM) and structural mean models (SMM) that are fitted using g-estimation.

In the linear and multiplicative (log-linear) cases, the two approaches lead to the same estimating equations [Clarke et al., 2011]. In the logistic case, they differ somewhat. This code is adapted from Clarke et al. [2011].

```
## A. Continuous outcome -- GMM and SMM approaches

# 1. Compact format
library(gmm)  # note that the gmm package also have a function called tsls
              # this document uses the sem package version of that command
asmm <- gmm(y ~ x, x=g)
print(cbind(coef(asmm),confint(asmm)))

# 2. Function format (assuming 2 genetic variables: g1 and g2)
asmmMoments <- function(theta,x){
 Y <- x[,1]
 X <- x[,2]
 Z1 <- x[,3]
 Z2 <- x[,4]
 # moments
 m1 <- (Y - theta[1] - theta[2]*X)
 m2 <- (Y - theta[1] - theta[2]*X)*Z1
 m3 <- (Y - theta[1] - theta[2]*X)*Z2
 return(cbind(m1,m2,m3))
}

four <- cbind(y,x,g1,g2)
asmm2 <- gmm(asmmMoments, x=four, t0=c(0,0))
print(cbind(coef(asmm2),confint(asmm2))) # Theta[2] is causal estimate

# 3. Stata code
gmm (y - {ey0} - x*{psi}), instruments(g1 g2)

ivregress gmm y (x = g1 g2)

ivreg2 y (x = g1 g2), gmm

## B. Binary outcome, multiplicative (log-linear) model -- GMM and SMM
    approaches

# 1. R code
```

```
msmmMoments <- function(theta,x){
 Y <- x[,1]
 X <- x[,2]
 Z1 <- x[,3]
 Z2 <- x[,4]
 m1 <- (Y*exp(-theta[1] - X*theta[2]) - 1)
 m2 <- (Y*exp(-theta[1] - X*theta[2]) - 1)*Z1
 m3 <- (Y*exp(-theta[1] - X*theta[2]) - 1)*Z2
 return(cbind(m1,m2,m3))
}
four = cbind(y,x,g1,g2)
msmm <- gmm(msmmMoments, x=four, t0=c(0,0))
print(cbind(coef(msmm), confint(msmm)))

# 2. Stata code
gmm (y*exp(-{ey0} - x*{psi}) - 1), instruments(g1 g2)

## C. Binary outcome, log-logistic model

# 1. SMM specification - R code
four <- cbind(y,x,g1,g2)

am <- glm(y ~ x*g1 + x*g2, family=binomial)
amfit <- coef(am)
linpred <- qlogis(fitted.values(am))

smMoments <- function(theta,x){
 # extract variables from x
 X <- x[,2]
 Z1 <- x[,3]
 Z2 <- x[,4]
 # moments
 s1 <- (plogis(linpred - theta[2]*X) - theta[1])
 s2 <- (plogis(linpred - theta[2]*X) - theta[1])*Z1
 s3 <- (plogis(linpred - theta[2]*X) - theta[1])*Z2
 return(cbind(s1,s2,s3))
}
sm <- gmm(smMoments, x=four, t0=c(0,0))
smfit <- coef(sm)
lsmmMoments <- function(theta,x){
 # extract variables from x
 Y <- x[,1]
 X <- x[,2]
 Z1 <- x[,3]
 Z2 <- x[,4]
 XZ1 <- X*Z1
 XZ2 <- X*Z2
 # association model moments
```

```
 linpred <- theta[1] + theta[2]*X + theta[3]*Z1 + theta[4]*Z2 +
    theta[5]*XZ1 + theta[6]*XZ2
 a1 <- (Y - plogis(linpred))
 a2 <- (Y - plogis(linpred))*X
 a3 <- (Y - plogis(linpred))*Z1
 a4 <- (Y - plogis(linpred))*Z2
 a5 <- (Y - plogis(linpred))*XZ1
 a6 <- (Y - plogis(linpred))*XZ2
 # structural model moments
 s1 <- (plogis(linpred - theta[8]*X) - theta[7])
 s2 <- (plogis(linpred - theta[8]*X) - theta[7])*Z1
 s3 <- (plogis(linpred - theta[8]*X) - theta[7])*Z2
 return(cbind(a1,a2,a3,a4,a5,a6,s1,s2,s3))
}
lsmm <- gmm(lsmmMoments, x=four, t0=c(amfit,smfit))
print(summary(lsmm))
print(cbind(coef(lsmm), confint(lsmm)))

# 2. SMM specification - Stata code
logit y x g1 g2 xg1 xg2
matrix from = e(b)
predict xblog, xb
gmm (invlogit(xblog - x*{psi}) - {ey0}), instruments(g1 g2) twostep

matrix from = (from,e(b))
* SEs incorrect here
gmm (y - invlogit({logit:x g1 g2 xg1 xg2} + {logitconst})) ///
(invlogit({logit:} + {logitconst} - x*{psi}) - {ey0}), ///
instruments(1:x g1 g2 xg1 xg2) instruments(2:g1 g2) ///
winitial(unadjusted, independent) from(from)

# 3. GMM specification - Stata code
gmm (y*invlogit(-{ey0} - x*{psi}) - 1), instruments(g1 g2)

gmm (y - invlogit({beta0} + x*{beta1})), instruments(g1 g2)
```

The SMM specification is due to Vansteelandt and Goetghebeur [2003].

3 Mendelian randomization analysis with summarized data

This section on Mendelian randomization methods with summarized data covers the inverse-variance weighted method that is equivalent asymptotically to the two-stage least squares method, as well as robust methods such as Egger regression, median-based estimation (including simple median and weighted median methods), a test for heterogeneity of the causal estimates from different genetic variants, and some code for the presentation of summarized data.

We assume that data are available on genetic associations with the risk factor and with the outcome in the form of beta-coefficients (`bx` and `by`) and standard errors (`bxse` and `byse`).

A one-sample Mendelian randomization setting is one in which data on genetic associations with the risk factor and with the outcome are estimated in the same individuals. A two-sample Mendelian randomization setting is one in which data on genetic associations with the risk factor and with the outcome are estimated in different samples [Pierce and Burgess, 2013]. Although two-sample investigations are not limited to those using summarized data, the use of summarized (particularly published) data means that genetic associations are often taken from separate sources, which may be non-overlapping [Burgess et al., 2015d].

3.1 Inverse-variance weighted method

The inverse-variance weighted method can be motivated in several ways [Burgess et al., 2013]. 1) It is a weighted mean of the ratio causal estimates from multiple genetic variants, where the weights are the inverse-variance weights also used in meta-analysis. 2) It is asymptotically equivalent to a two-stage least squares analysis (and so can be motivated using an allele score). 3) It is the coefficient from a weighted linear regression of the gene–outcome coefficients on the gene-risk factor coefficients (intercept is constrained to equal zero).

We initially present the inverse-variance weighted estimate and standard error as it was initially proposed (using a fixed-effect assumption – that all the genetic variants identify the same causal effect, and using a simple formulation of the variances used as weights) [Ehret and others, 2011]. We then give random-effects estimates, as well as a correction to the variances that should be used in a one-sample setting, particularly when the instruments are weak. We initially assume that genetic variants are uncorrelated (not in linkage disequilibrium).

```
## A. Genetic variants uncorrelated, no heterogeneity -- fixed-effect model

# 1. Gene score formula:
betafirst.fixed = sum(bx*by/byse^2)/sum(bx^2/byse^2)
  sefirst.fixed = sqrt(1/sum(bx^2/byse^2))

# 2. Meta-analysis:
library(meta)
betafirst.fixed = metagen(by/bx, byse/bx)$TE.fixed
  sefirst.fixed = metagen(by/bx, byse/bx)$seTE.fixed

# 3. Weighted linear regression:
betafirst.fixed = lm(by~bx-1, weights=byse^-2)$coef[1]
  sefirst.fixed = summary(lm(by~bx-1,
      weights=byse^-2))$coef[1,2]/summary(lm(by~bx-1,
      weights=byse^-2))$sigma

## B. Genetic variants uncorrelated, heterogeneity -- random-effects model
#     (the use of a random-effects model is preferred,
#       particularly if there is any chance of heterogeneity
#       in the estimates obtained using different genetic variants)

# 1. Additive random-effects model
betafirst.addran = metagen(by/bx, abs(byse/bx))$TE.random
  sefirst.addran = metagen(by/bx, abs(byse/bx))$seTE.random

# 2. Multiplicative random-effects model
       reg.first = summary(lm(by~bx-1, weights=byse^-2))
betafirst.mulran = reg.first$coef[1]
  sefirst.mulran = reg.first$coef[1,2]/min(reg.first$sigma,1)
```

A random-effects model (either additive or multiplicative) should be used for combining causal estimates from multiple genetic variants. If there is no heterogeneity, then there is no loss, as the estimate from a random-effects analysis equals that from a fixed-effect analysis. If there is heterogeneity, then the estimate from a fixed-effect analysis will be too precise.

```
## C. Genetic variants uncorrelated, heterogeneity -- correction for
    one-sample setting or sample overlap

# 1. Additive random-effects model
betasecond.addran = metagen(by/bx,
    sqrt(byse^2/bx^2+by^2*bxse^2/bx^4))$TE.random
  sesecond.addran = metagen(by/bx,
      sqrt(byse^2/bx^2+by^2*bxse^2/bx^4))$seTE.random

# 2. Multiplicative random-effects model
       reg.second = summary(lm(by~bx-1,
           weights=(byse^2+by^2*bxse^2/bx^2)^-1))

betasecond.mulran = reg.second$coef[1]
  sesecond.mulran = reg.second$coef[1,2]/min(reg.second$sigma,1)

# 3. Incorporating correlation term -- additive random-effects model
theta = 0.1                          # correlation term

betasecond.theta.addran = metagen(by/bx,
 sqrt(byse^2/bx^2+by^2*bxse^2/bx^4
 -2*theta*by*bxse*byse/bx^3))$TE.random

  sesecond.theta.addran = metagen(by/bx,
 sqrt(byse^2/bx^2+by^2*bxse^2/bx^4
 -2*theta*by*bxse*byse/bx^3))$seTE.random

# 4. Incorporating correlation term -- multiplicative random-effects model
theta = 0.1                          # correlation term

reg.second.theta = summary(lm(by~bx-1,
 weights=(byse^2+by^2*bxse^2/bx^2-2*theta*by*bxse*byse/bx)^-1))

betasecond.theta.mulran = reg.second.theta$coef[1]
  sesecond.theta.mulran =
      reg.second.theta$coef[1,2]/min(reg.second.theta$sigma,1)
```

When the genetic associations with the risk factor and with the outcome are estimated in non-overlapping datasets (a two-sample setting), the estimates in section B give reasonable coverage rates and those in section C give overcoverage (confidence intervals are too wide) [unpublished]. However, when the datasets overlap (as is often the case in practice for major genetic consortia), the coverage of estimates in section

B is below nominal levels and Type 1 error rates are too frequent. This is particularly true when the genetic variants are weak. In the one-sample or overlapping sample setting, estimates in section C have much better coverage properties.

The correlation term θ is the correlation between the genetic associations with the risk factor and the genetic associations with the outcome. This is zero in a two-sample setting, and approximately the same as the correlation between the risk factor and the outcome in a one-sample setting.

```
## D. Genetic variants correlated

Omega = byse%o%byse*rho
betafirst.fixed.correl =
    solve(t(bx)%*%solve(Omega)%*%bx)*t(bx)%*%solve(Omega)%*%by
sefirst.fixed.correl = sqrt(solve(t(bx)%*%solve(Omega)%*%bx))
```

When genetic variants are correlated, the causal effect can be estimated using generalized weighted linear regression [Burgess et al., 2015b]. The matrix `rho` is the matrix of correlations between the genetic variants. This method can give odd results, particularly if the genetic variants are highly correlated, or the correlations are misspecified.

3.2 Egger regression

Egger regression is an extension to the inverse-variance weighted method to allow genetic variants to have pleiotropic (direct) effects on the outcome that do not operate via the risk factor [Bowden et al., 2015a]. The genetic variants are allowed to violate the instrumental variable assumptions, and specifically the exclusion restriction assumption. As the genetic variants are not instrumental variables, the assumption of homogeneity of causal estimates from each genetic variant can no longer be made.

In Egger regression, a weighted regression of the gene–outcome coefficients on the gene-risk factor coefficients is performed, except that the intercept term is estimated as part of the model. If the estimated intercept differs from zero, this provides evidence that there is directional (or unbalanced) pleiotropy. This means that the pleiotropic effects of the genetic variants do not average to have mean zero.

Under an additional assumption (the InSIDE – instrument strength independent of direct effect – assumption), the slope coefficient from Egger regression is a consistent estimate of the causal effect even the presence of directional pleiotropy. The InSIDE assumption states that the distribution of pleiotropic effects across the genetic variants is independent of the distribution of genetic associations with the risk factor. For a finite number of genetic variants, the estimate is consistent if the direct effects are uncorrelated with the associations with the risk factor (the instrument strength).

```
## A. Genetic variants uncorrelated, multiplicative random-effects model

# 1. Test for directional pleiotropy
inter_Egger  = summary(lm(by~bx, weights=byse^-2))$coef[1,1]
seinter_Egger = summary(lm(by~bx, weights=byse^-2))$coef[1,2]/
           min(summary(lm(by~bx, weights=byse^-2))$sigma, 1)

# 2. Estimate of causal effect (under InSIDE assumption)
beta_Egger   = summary(lm(by~bx, weights=byse^-2))$coef[2,1]
sebeta_Egger = summary(lm(by~bx, weights=byse^-2))$coef[2,2]/
           min(summary(lm(by~bx, weights=byse^-2))$sigma, 1)
```

3.3 Median-based estimation

The inverse-variance weighted estimate is a weighted mean of the estimates from each individual genetic variant. This has a 0% breakdown limit, as if one of the genetic variants is not a valid instrumental variable (and so the estimate from that variant is biased), then the inverse-variance weighted estimate will be inconsistent. On the contrary, the median has a 50% breakdown limit, and so the median of the estimates from each individual genetic variant will be a consistent estimate of the causal effect even if up to 50% of the genetic variants are not valid instrumental variables [Han, 2008].

Alternatively, a weighted median can be calculated by considering as the median of a probability distribution in which the probability of each causal estimate is proportional to a given weight. In particular, we use the same inverse-variance weights as in the inverse-variance weighted method. The consistency condition is now that genetic variants representing at least 50% of the weight are valid instrumental variables [Bowden et al., 2015b].

```
## A. Median-based estimation

weighted.median <- function(betaIV.in, weights.in) {
  betaIV.order = betaIV.in[order(betaIV.in)]
  weights.order = weights.in[order(betaIV.in)]
  weights.sum  = cumsum(weights.order)-0.5*weights.order
  weights.sum  = weights.sum/sum(weights.order)
       below   = max(which(weights.sum<0.5))
  weighted.est = betaIV.order[below] +
      (betaIV.order[below+1]-betaIV.order[below])*
                 (0.5-weights.sum[below])/(weights.sum[below+1]-weights.sum[below])
  return(weighted.est) }

weighted.median.boot = function(betaXG.in, betaYG.in, sebetaXG.in,
    sebetaYG.in, weights.in){
med = NULL
for(i in 1:1000){
 betaXG.boot = rnorm(length(betaXG.in), mean=betaXG.in, sd=sebetaXG.in)
 betaYG.boot = rnorm(length(betaYG.in), mean=betaYG.in, sd=sebetaYG.in)
 betaIV.boot = betaYG.boot/betaXG.boot
 med[i] = weighted.median(betaIV.boot, weights.in)
 }
return(sd(med)) }

# 1. Simple median estimator
betaIV           = by/bx
weights          = rep(1, length(bx))
betaSIMPLEMED    = weighted.median(betaIV, weights)
sebetaSIMPLEMED  = weighted.median.boot(bx, by, bxse, byse, weights)

# 2. Weighted median estimator
```

```
betaIV           = by/bx
weights          = (byse/bx)^-2
betaWEIGHTEDMED  = weighted.median(betaIV, weights)
sebetaWEIGHTEDMED = weighted.median.boot(bx, by, bxse, byse, weights)
```

The standard error is estimated using a parametric bootstrap.

3.4 Heterogeneity test and presentation of data

A heterogeneity test can be performed using Cochran's Q statistic [Greco et al., 2015]. A significant test statistic corresponds to rejection of the null hypothesis that all genetic variants identify the same causal effect. Substantial heterogeneity is an indication of potential violation of the instrumental variable assumptions for one or more genetic variants.

```
metagen(by/bx, byse/bx)
p.hetero = 1-pchisq(metagen(by/bx, byse/bx)$Q,
   metagen(by/bx, byse/bx)$df.Q)
```

Heterogeneity and directional pleiotropy can also be detects visual using a scatter plot or a funnel plot based on the associations of the individual genetic variants.

```
# 1. Scatter plot (gene--outcome associations against gene--risk factor
    associations)
#               (lines represent 95\% confidence intervals)
by = by*sign(bx); bx = abs(bx)
plot(bx, by, xlim=c(min(bx-2*bxse, 0), max(bx+2*bxse)),
 ylim=c(min(by-2*byse, 0), max(by+2*byse, 0)))
for (j in 1:length(bx)) {
 lines(c(bx[j],bx[j]), c(by[j]-1.96*byse[j], by[j]+1.96*byse[j]))
 lines(c(bx[j]-1.96*bxse[j],bx[j]+1.96*bxse[j]), c(by[j], by[j]))
 }
abline(h=0, lwd=1); abline(v=0, lwd=1)

# 2. Funnel plot (measure of precision of IV estimate -- specifically the
    reciprocal of the standard error
#               versus the IV estimate)
plot(by/bx, bx/byse, xlim=c(min((by-2*byse)/bx),
  max((by+2*byse)/bx)), ylim=c(0, max(bx/byse)))
for (j in 1:length(bx)) {
 lines(c((by[j]-1.96*byse[j])/bx[j], (by[j]+1.96*byse[j])/bx[j]),
  c(bx[j]/byse[j], bx[j]/byse[j]))
 }
abline(h=0, lwd=1); abline(v=0, lwd=1)
```

4 Additional analyses

4.1 Multivariable Mendelian randomization

In multivariable Mendelian randomization, genetic variants are allowed to affect more than one risk factor, but they have to satisfy the instrumental variable assumptions for the set of risk factors. So, genetic variants are allowed to have pleiotropic effects, provided that the pleiotropic effects are limited to the set of risk factors under investigation [Burgess and Thompson, 2015a]. The effects of each of the risk factors are all estimated in one model.

With individual-level data, a two-stage method can be used, except that the first-stage model is now a multivariate regression model with each of the risk factors as a dependent variable. With summarized data, an extension to the inverse-variance weighted method can be performed. Instead of a univariable regression model, the gene–outcome associations can be regressed on all the gene–risk factor associations simultaneously in a multivariable regression model.

```
## A. Individual-level data -- assuming two risk factors

library(sem)
beta1_tsls = tsls(y, cbind(x1, x2, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$coef[1]
beta2_tsls = tsls(y, cbind(x1, x2, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$coef[2]
se1_tsls  = sqrt(tsls(y, cbind(x1, x2, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$V[1,1])
se2_tsls  = sqrt(tsls(y, cbind(x1, x2, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$V[2,2])

library(ivpack)
ivmodel = ivreg(y~cbind(x1,x2)|g, x=TRUE)
summary(ivmodel)
```

With summarized data, the inverse-variance weighted method can be used, except that instead of a univariate weighted linear regression model, a multivariate weighted linear regression model can be employed [Burgess et al., 2015c].

```
## B. Summarized data

beta_multivar = summary(lm(by~bx1+bx2-1, weights=byse^-2))$coef[1,1]
se_multivar  = summary(lm(by~bx1+bx2-1, weights=byse^-2))$coef[1,2]/
          min(summary(lm(by~bx1+bx2-1, weights=byse^-2))$sigma, 1)
```

4.2 Non-linear estimation

If the relationship between the risk factor and the outcome is non-linear, then an estimate that ignores this non-linearity can be interpreted as a population-averaged effect of an intervention in the distribution of the risk factor in the population. Alternatively, separate causal estimates can be obtained in strata of the population. These estimates can be used to determine the shape of the risk factor–outcome relationship. However, if the risk factor is stratified on directly, then associations between the genetic variants and the outcome will be distorted, as the risk factor is on the causal pathway from the genetic variants to the outcome [Burgess et al., 2014].

A solution to this is to first regress the risk factor on the genetic variants, and then to stratify on the residuals from this regression. The residual represents the non-genetic component of the risk factor, or the expected value of the exposure if the genetic variants took the value zero. The causal estimate (the localized average causal effect) can then be obtained in each stratum. The assumption is made that the effect of the genetic variants on the risk factor is equal in all individuals.

```
## A. Estimation of localized average causal effect in 5 strata

x0 = lm(x~g)$residuals
strata =
    (order(x0)>N/5)+(order(x0)>2*N/5)+(order(x0)>3*N/5)+(order(x0)>4*N/5)

beta_lace = NULL; se_lace = NULL
for (j in 0:4) {
 beta_lace[j+1] = lm(y[strata==j]~g[strata==j])$coef[2]/lm(x~g)$coef[2]
   se_lace[j+1] =
       summary(lm(y[strata==j]~g[strata==j]))$coef[2,2]/lm(x~g)$coef[2]
              }
```

References

Anderson, T. and Rubin, H. 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics*, 21(1):570–582. (page 6).

Bowden, J., Davey Smith, G., and Burgess, S. 2015a. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525. (page 18).

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. 2015b. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Available at https://www.academia.edu/15479132/Consistent. (page 19).

Bowden, J. and Vansteelandt, S. 2011. Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine*, 30(6):678–694. (page 7).

Buonaccorsi, J. 2005. *Encyclopedia of Biostatistics*, chapter Fieller's theorem, pages 1951–1952. Wiley. (page 6).

Burgess, S., Butterworth, A., and Thompson, S. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665. (page 15).

Burgess, S., Butterworth, A. S., and Thompson, J. R. 2015a. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors. *Journal of Clinical Epidemiology*. (page 3).

Burgess, S. and CHD CRP Genetics Collaboration 2013. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32(27):4726–4747. (page 8).

Burgess, S., Davies, N., Thompson, S., and EPIC-InterAct Consortium 2014. Instrumental variable analysis with a nonlinear exposure–outcome relationship. *Epidemiology*, 25(6):877–885. (page 23).

Burgess, S., Dudbridge, F., and Thompson, S. 2015b. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. Available at https://www.academia.edu/15479109/Combining. (page 17).

Burgess, S., Dudbridge, F., and Thompson, S. G. 2015c. Re: "Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects". *American Journal of Epidemiology*, 181(4):290–291. (page 22).

Burgess, S., Scott, R., Timpson, N., Davey Smith, G., Thompson, S., and EPIC-InterAct Consortium 2015d. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology*, 30(7):543–552. (page 14).

Burgess, S., Small, D. S., and Thompson, S. G. 2015e. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*. (page 4).

Burgess, S. and Thompson, S. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4):1134–1144. (page 5).

Burgess, S. and Thompson, S. 2015a. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, 181(4):251–260. (page 22).

Burgess, S. and Thompson, S. G. 2015b. *Mendelian randomization: methods for using genetic variants in causal estimation*. Chapman & Hall. (page 4).

Cai, B., Small, D., and Ten Have, T. 2011. Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. *Statistics in Medicine*, 30(15):1809–1824. (page 8).

Clarke, P., Palmer, T., and Windmeijer, F. 2011. Estimating structural mean models with multiple instrumental variables using the generalised method of moments. The Centre for Market and Public Organisation 11/266, Centre for Market and Public Organisation, University of Bristol, UK. (page 11).

Clarke, P. S. and Windmeijer, F. 2012. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652. (page 7).

Didelez, V. and Sheehan, N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330. (page 6).

Ehret, G., Munroe, P., Rice, K., et al. (The International Consortium for Blood Pressure Genome-Wide Association Studies) 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478:103–109. (page 15).

Fieller, E. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 16(2):175–185. (page 6).

Glymour, M., Tchetgen Tchetgen, E., and Robins, J. 2012. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339. (page 3).

Greco, M., Minelli, C., Sheehan, N. A., and Thompson, J. R. 2015. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine*, 34(21):2926–2940. (page 21).

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729. (page 5).

Greenland, S., Robins, J., and Pearl, J. 1999. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46. (page 7).

Han, C. 2008. Detecting invalid instruments using L1-GMM. *Economics Letters*, 101:285–287. (page 19).

Hernán, M. and Robins, J. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372. (page 3).

Lawlor, D., Harbord, R., Sterne, J., Timpson, N., and Davey Smith, G. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163. (page 5).

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. 2006. Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267. (page 5).

Mikusheva, A. 2010. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247. (page 6).

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdorff, M. 2000. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19(14):1849–1864. (page 9).

Palmer, T., Sterne, J., Harbord, R., et al. 2011. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *American Journal of Epidemiology*, 173(12):1392–1403. (page 7).

Palmer, T., Thompson, J., Tobin, M., Sheehan, N., and Burton, P. 2008. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, 37(5):1161–1168. (page 9).

Pierce, B. and Burgess, S. 2013. Efficient design for Mendelian randomization studies: subsample and two-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184. (page 14).

Terza, J., Basu, A., and Rathouz, P. 2008. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543. (page 9).

VanderWeele, T., Tchetgen Tchetgen, E., Cornelis, M., and Kraft, P. 2014. Methodological challenges in Mendelian randomization. *Epidemiology*, 25(3):427–435. (page 3).

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. 2011. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422. (pages 7, 8, 9).

Vansteelandt, S. and Goetghebeur, E. 2003. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):817–835. (page 13).

Wald, A. 1940. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11(3):284–300. (page 5).